DEIM Forum 2012 F5-2

カテゴリ階層を考慮した確率的トピックモデルのモデル選択付き学習

山本　浩平†　　江口　浩二†　　高須　淳宏††

† 神戸大学　〒 657–8501 神戸市灘区六甲台町 1-1
†† 国立情報学研究所　〒 101–8430 東京都千代田区一ツ橋 2-1-2
E-mail: †yamamoto@cs25.scitec.kobe-u.ac.jp, eguchi@port.kobe-u.ac.jp, ††takasu@nii.ac.jp

あらまし　情報の継続的な増加に伴い，大規模な情報にアクセスする有効な手段の一つとして，文書に階層カテゴリ情報を自動的に付与することによる，文書集合のインデックス化と階層化が望まれている．近年注目されつつある確率的トピックモデルが有効な手段として考えられるが，その代表的なモデルである LDA (Latent Dirichlet Allocation) ではカテゴリ情報を明示的にモデル化しないため，新たなモデル化が求められる．そこで，我々はカテゴリ階層構造を持つ文書集合に適したトピックモデルとして DirTM (Directory Topic Model) を提案し，ギブスサンプリングに基づくモデル選択付き学習によってモデルパラメータを推定する．

キーワード　トピックモデル，潜在的ディリクレ配分法，ギブスサンプリング，情報量基準

Kohei YAMAMOTO†, Koji EGUCHI†, and Atsuhiro TAKASU††

† Kobe University
1-1 Rokkodaicho, Nada, Kobe 657–8501 Japan
†† National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda, Tokyo 101–8430 Japan
E-mail: †yamamoto@cs25.scitec.kobe-u.ac.jp, eguchi@port.kobe-u.ac.jp, ††takasu@nii.ac.jp

1. はじめに

近年，情報技術の発展により，世に存在する様々な情報の規模が爆発的に増大している．現在，大規模な情報の潜在的構造を抽出するための手法として，潜在的ディリクレ配分法 (LDA: Latent Dirichlet Allocation) に代表されるトピックモデルが存在する [5]．トピックモデルとは，文書はある特徴を持った単語の確率分布である潜在トピックの混合分布から生成される，と仮定した生成モデルである．与えられた文書から，トピックモデルを用いて潜在トピックを推定することによって，情報検索，情報推薦，文書分類などへの適用が可能となる．トピックモデルは，テキストデータ [13]，ネットワークデータ [1]，画像データ [7] など，様々なデータに対して適用されている．

ところで，近年，ユーザの求める情報へのアクセスを容易にする目的で，テキストデータに対し木構造に代表されるカテゴリ階層構造を付与して提供されることが少なくない．カテゴリは文書の大まかな内容を表しており，階層の上位に位置するカテゴリは抽象的な内容を，下位に位置するカテゴリは具体的な内容を表していると考えられる．つまり，階層の上位に位置するカテゴリが付与されている文書は，一般的な内容であり，多様な潜在トピックが含まれていると考えられる．以上に述べたことから，カテゴリ階層構造は文書集合に対して潜在トピックを推定する際に有用であると考えられる．

しかし，代表的なトピックモデルである LDA では，文書に付与されているカテゴリをモデル化の際に考慮に入れていない．そこで，カテゴリ階層構造を持つ文書集合に対して適するトピックモデルとして，DirTM (Directory Topic Model) を提案する．DirTM では，木構造をなしているカテゴリ階層構造において，内部ノードであるカテゴリに属する文書は，葉ノードであるカテゴリに属する文書のトピック分布の混合分布から生成されると仮定する．そして，葉ノードに属する文書のトピック分布は，LDA の仮定に基づいて生成されるとする．また，各内部ノードごとに，多項分布によって子ノードへの辺を生成し，その多項分布の事前分布としてディリクレ事前分布を用いる．それぞれの内部ノードから葉ノードに至るパスについて，そのパスを構成する辺の同時確率により葉ノードに属する各文書の文書-トピック分布の期待値を重み付けし，内部ノードにおいてそれらの文書-トピック分布の期待値が混合される．また，それぞれの葉ノードにおける文書-トピック分布のディリクレ事前分布を決定するハイパーパラメータを推定し，葉ノードであるカテゴリごとに最適な潜在トピックが割り当てられるようにする．

このとき，葉ノードに属する文書の集合の規模が十分大きく

ない場合に，ハイパーパラメータの過適合が起こるという問題が考えられる．よって，葉ノードにおいて正確な文書-トピック分布の推定が行われない可能性がある．そこで，本研究では，DirTM を推定する際に，情報量基準を用いて最適なモデルの選択を行う．すなわち，木構造をなしているカテゴリ階層構造の部分木に対して，部分木中の全てのノードを併合したときとそうでないときで，情報量基準の値を比較する．もし，部分木中の全てのノードを併合したときのほうがモデルとして適切であれば，モデルを更新し，部分木中の全てのノードに属する文書について，トピック分布を推定し直す．この操作を部分木の根をなす内部ノード全てに対して行い，先に述べた過適合の問題を回避することを目指す．

本研究では，階層カテゴリ構造を付与された文書集合に対して適当なトピックモデルである DirTM を提案し，その未知パラメータを推定する過程において，情報量規準を用いたモデル選択を行う．そして，訓練文書に対して提案手法，また LDA を用いることでモデルを推定し，テストセット対数尤度を測定することで精度の比較を行い，提案手法の有効性を示す．

2. 関 連 研 究

まず，提案手法の基礎となっている研究として，LDA とその他の関連するトピックモデルについてその概要を説明する．

2.1 LDA

潜在的ディリクレ配分法 (LDA: Latent Dirichlet Allocation) は，代表的なトピックモデルである．LDA は，文書はある特徴を持った単語の分布であるトピックの混合分布から生成されると仮定する生成モデルである．ここで，文書-トピック多項分布とトピック-単語多項分布のそれぞれについて，ディリクレ事前分布を仮定する．Blei らによって最初に LDA が提案された論文 [5] では，モデルの推定方法として変分ベイズ法が用いられていたが，本研究では，実装が比較的容易で大規模なコーパスに対して他の手法より効率的である，ギブスサンプリングを推定手法として用いている．この手法は Griffiths らによって提案された [8]．以下に LDA の文書生成過程を示す．ここで，$\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ はそれぞれ文書-トピック多項分布とトピック-単語多項分布に対するディリクレ事前分布のパラメータ (ハイパーパラメータ) である．$\boldsymbol{\alpha}$ はトピック数分，$\boldsymbol{\beta}$ は語彙数 (異なり語数) 分の成分を持つベクトルで表される．$\boldsymbol{w}_{in}$ は文書 i の n 番目の単語に関する指示ベクトル (indicator vector) で表される(注1)．$\boldsymbol{z}_{in}$ は文書 i の n 番目の単語に割り当たっているトピックに関する指示ベクトルで表される．D は文書数，T はトピック数，N_i は文書 i の文書長を示す．

（1） ディリクレ分布 Dir($\boldsymbol{\alpha}$) から，各文書 $i \in \{1, \cdots, D\}$ に対して，多項分布パラメータ $\boldsymbol{\theta}_i$ を選ぶ．

（2） ディリクレ分布 Dir($\boldsymbol{\beta}$) から，各トピック $t \in \{1, \cdots, T\}$ に対して，多項分布パラメータ $\boldsymbol{\phi}_t$ を選ぶ．

（3） 文書 i 内の各語 $\boldsymbol{w}_{in} (n \in \{1, \cdots, N_i\})$ に対して，

（a） 多項分布 Mult($\boldsymbol{\theta}_i$) からトピック $\boldsymbol{z}_{in}$ を選ぶ．

（b） 多項分布 Mult($\boldsymbol{\phi}_t$) から単語 $\boldsymbol{w}_{in}$ を選ぶ．

ディリクレ分布 Dir($\boldsymbol{\alpha}$) において，ハイパーパラメータのベクトル $\boldsymbol{\alpha}$，$\boldsymbol{\beta}$ の各成分は，元々は α_t $(t \in \{1, \cdots, T\})$ は $\alpha_t = 50/T$，β_w $(w \in \{1, \cdots, W\}$, W は文書集合の語彙数) は $\beta_w = 0.1$ とする対称ディリクレ分布が経験的に用いられていた [8]．その後，$\boldsymbol{\alpha}$ の各成分については，不動点反復による推定を行うことによる非対称ディリクレ分布を用い，他方，$\boldsymbol{\beta}$ の各成分は全て同じとする対称ディリクレ分布を用いることで，モデル推定の精度が改善されることが確認されている [11], [15]．

2.2 その他の関連するトピックモデル

階層構造を導入したトピックモデルについて概説する．無閉路有向グラフ (DAG: Directed acyclic graph) で表される潜在トピックの階層構造を推測するトピックモデルである，Pachinko Allocation Model (PAM) が存在する [10]．これは Li らによって提案されている．PAM において，文書集合から教師なし学習によってトピック間の階層関係は推測され，潜在トピックの推定に用いられる．一方，DirTM では既知のカテゴリ階層が付与された文書集合に対して，文書集合における各文書がどのカテゴリに属するか，という情報を用いてトピック分布を推定する．人手で付与されたカテゴリ階層構造を持つような文書集合が近年多く提供されており，そのような文書集合を対象としている点で，PAM と DirTM は目的が異なる．

評価，人気度などの連続値，またはカテゴリなどの離散値が各文書に付与された文書集合を対象に，潜在トピックを推定する Supervised Latent Dirichlet Allocation (sLDA) が，Blei らによって提案されている [4]．また，カテゴリ付き文書集合を対象にしたトピックモデルとしては，Discriminative Latent Dirichlet Allocation (DiscLDA) が Lacoste-Julien らによって提案されている [9]．他には，潜在トピックとユーザなどによって文書に付与されるタグを 1 対 1 に対応付けて利用する，Ramage らによって提案された Labeled LDA が存在する [12]．sLDA, DiscLDA，Labeled LDA により，各文書に非階層カテゴリが付与された文書集合を対象にして，潜在トピックを効果的に推定することができるが，いずれのモデルも階層カテゴリを想定したものではない．これに対して，DirTM ではカテゴリ階層構造を付与された文書集合を対象にしている点で目的が異なる．

以上のように，本研究で提案する DirTM は，カテゴリ階層が既に付与された文書集合を対象としている点，カテゴリ階層の意味的な依存関係を考慮してモデル推定に活かしている点で，従来の研究とは目的が異なる．

(注1)：この場合，語彙数だけの成分を持つベクトルであり，文書 i の n 番目の単語にあたる語彙に対応する成分が 1，それ以外の成分が 0 であるようなベクトルである．なお，この後に出てくる指示ベクトルはトピック数だけの成分を持つベクトルである．

3. 階層カテゴリ構造を考慮した確率的トピックモデル

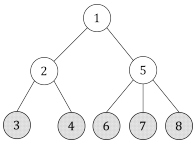

図 1　A example of category tree

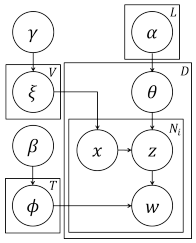

図 2　A graphical model of DirTM

カテゴリが付与された文書からなる文書集合を考える．このような文書集合に付与されるカテゴリは一般に図 1 のように木構造をしており，木の各カテゴリノードそれぞれに関連する文書が割り当てられている．カテゴリは，それに属する文書の大まかな内容を表している．つまり，同じカテゴリに属する文書の内容は類似している可能性が高く，類似したトピック分布を持つ一方で，異なったカテゴリに属す文書とは内容もトピック分布も異なると考えられる．また，階層の上位のカテゴリは抽象的な内容を，下位のカテゴリは具体的な内容を表している．つまり，上位のカテゴリが付与された文書は，一般的な内容を持ち潜在トピックが多様であり，下位のカテゴリが付与された文書は，より具体的な内容であるといえる．

代表的なトピックモデルである LDA では，モデル推定の際に先に述べたようなカテゴリを考慮に入れておらず，その点で拡張の余地があるといえる．そこで，カテゴリ階層構造を持つ文書集合に適したトピックモデルとして，DirTM (Directory Topic Model) を提案する．

DirTM では，カテゴリ木の葉ノードに属する文書の文書-トピック分布の混合分布から，内部ノードに属する文書が生成されると仮定している．そして，葉ノードに属する文書の文書-トピック分布は，それぞれの葉ノードであるカテゴリのハイパーパラメータ $\boldsymbol{\alpha}$ を用いた LDA の仮定に基づいて生成されるとする．葉ノードの文書-トピック分布の混合は以下のようにして行われる．まず，各内部ノードごとに，多項分布によって子ノードへの辺を生成し，その多項分布の事前分布としてディリクレ事前分布を用いる．それぞれの内部ノードから葉ノードに至るパスについて，そのパスを構成する辺の同時確率により葉ノードの文書-トピック分布の期待値を重み付けし，内部ノードにおいてそれらの文書-トピック分布の期待値が混合される．また，それぞれの葉ノードにおける文書-トピック分布の非対称ディリクレ事前分布を決定するハイパーパラメータ $\boldsymbol{\alpha}$ を推定し，カテゴリごとに最適な潜在トピックが割り当てられるようにする．一方，トピック-単語多項分布に関しては，$\boldsymbol{\beta}$ を固定した対称ディリクレ分布を用い，文書の属すカテゴリとは独立としている．

具体的な例を用いて説明する．図 1 のようなカテゴリ木において，葉ノード 3，4，6，7，8 に属する文書の文書-トピック分布は，それぞれの葉ノードごとに異なるハイパーパラメータ $\boldsymbol{\alpha}_\ell$ によって推定される．このことは，それぞれの葉ノードによっての内容の違いが，それぞれ異なるハイパーパラメータによって明確にされているということである．また，内部ノード 2 に属する文書の文書-トピック分布は，葉ノード 3，4 に属する文書の文書-トピック分布の期待値がパスの生成確率によって重み付けされ，混合されている．つまり，内部ノード 2 に属する文書の内容はその子ノードの内容を含むより一般的なものになっているということになる．同様に考えると，内部ノード 5 はその子である葉ノード 6，7，8 の文書-トピック分布の混合分布，カテゴリ木の根ノード 1 は木における全ての葉ノードの文書-トピック分布の混合分布を持ち，カテゴリ階層の上位に位置するノードほど，一般的な内容を表すカテゴリになっているといえる．

以下で，葉ノードと内部ノードのそれぞれにおける文書の生成過程について説明する．

3.1 葉ノード

DirTM において，葉ノードでの文書のモデリングは，葉ノード ℓ によって異なるハイパーパラメータ $\boldsymbol{\alpha_\ell}$ を用いることを除いて，LDA と同様である．LDA では，ハイパーパラメータ $\boldsymbol{\alpha}$ の非対称ディリクレ事前分布から，文書-トピック多項分布のパラメータを選び，全ての文書について同じ多項分布で文書生成を行う．DirTM では，葉ノードごとに異なるハイパーパラメータ $\boldsymbol{\alpha_\ell}$ を用いることによって，葉ノードであるカテゴリごとの内容の違いを明確にする．

DirTM における，葉ノード ℓ に属する文書の生成過程を以下に示す．ここで，$\mathcal{D}_\ell$ は葉ノード ℓ に属する文書の集合を表す．

（1） ディリクレ分布 Dir($\boldsymbol{\alpha}_\ell$) から，各文書 $i \in \mathcal{D}_\ell$ に対して，多項分布パラメータ $\boldsymbol{\theta}_{\ell i}$ を選ぶ．

（2） ディリクレ分布 Dir($\boldsymbol{\beta}$) から，各トピック $t \in \{1, \cdots, T\}$ に対して，多項分布パラメータ $\boldsymbol{\phi}_t$ を選ぶ．

（3） 文書 i 内の各語 $\boldsymbol{w}_{in}$ ($n \in \{1, \cdots, N_i\}$) に対して，

（a） 多項分布 Mult($\boldsymbol{\theta}_{\ell i}$) からトピック $\boldsymbol{z}_{in}$ を選ぶ．

（b） 多項分布 Mult($\boldsymbol{\phi}_t$) から単語 $\boldsymbol{w}_{in}$ を選ぶ．

3.2 内部ノード

DirTM において，内部ノードに割り当てられた文書に関するモデリングについて考える．カテゴリ木を構成するノード集合に対して，完全グラフの各ノードについてループが存在するようなグラフを仮定する．つまり，カテゴリ木における内部ノードに属している文書 i を考えると，i が属するノード $v \in \{1, \cdots, V\}$ が任意の各ノード $u \in \{1, \cdots, U\}$ に隣接すると仮定する．ここで，V は内部ノード数，U は全ノード数を表す．各内部ノードごとに，ハイパーパラメータ $\boldsymbol{\gamma}$ のディリクレ事前分布から隣接ノードへの辺に関する多項分布パラメータが選ばれるとする．この多項分布において，先ほど仮定したグラフの部分木である，実際のカテゴリ木を構成する辺にのみ，高い確率が割り当てられる．また，パス確率 $x_{i\ell}$ を，先ほど仮定したグラフにおいて文書 i が属するノード v からカテゴリ木の葉ノードに対応するノード $\ell \in \mathcal{L}$ に至るパスであるとする．ここで，$\mathcal{L}$ は全葉ノードの集合である．また，それぞれの葉ノー

ドについて，属する文書のトピック分布の期待値をその葉ノードのトピック分布とする．これにより，その葉ノードの内容をおおまかに表したトピック分布が得られることになる．内部ノードに属する文書のトピック分布は葉ノードに関するトピック分布の混合分布で表されると仮定する．この際に，実際にカテゴリ木において v の子孫である葉ノードへのパス確率と比較して，そうでない葉ノードへのパス確率は極めて小さくなる．それらの各パス確率に応じた比率で，各葉ノードのトピック分布が混合され，それが文書 i のトピック分布となる．

DirTM における，内部ノード v に属する文書の生成過程を以下に示す．

（1） ディリクレ分布 $\mathrm{Dir}(\boldsymbol{\beta})$ から，各トピック $t \in \{1,\cdots,T\}$ に対して，多項分布パラメータ $\boldsymbol{\phi}_t$ を選ぶ．

（2） カテゴリ木の各内部ノードに対して，

（a） ディリクレ分布 $\mathrm{Dir}(\boldsymbol{\gamma})$ から，各内部ノード v に対して，多項分布パラメータ $\boldsymbol{\xi}_v$ を選ぶ．

（3） 内部ノード v に属する文書 i 内の各語 $\boldsymbol{w}_{in}$ $(n \in \{1,\cdots,N_i\})$ に対して，

（a） v が葉ノードになるまで，

i. 多項分布 $\mathrm{Mult}(\boldsymbol{\xi}_v)$ から，隣接ノード v' を選ぶ．

ii. v' を新たに v とし，$v \in \mathcal{L}$ であれば v を ℓ とする．

（b） ディリクレ分布 $\mathrm{Dir}(\boldsymbol{\alpha}_\ell)$ から，各文書 $j \in \mathcal{D}_\ell$ に対する多項分布パラメータ $\boldsymbol{\theta}_{\ell j}$ を選び，その期待値 $\boldsymbol{\theta}_\ell$ を得る．

（c） 多項分布 $\mathrm{Mult}(\boldsymbol{\theta}_\ell)$ からトピック $\boldsymbol{z}_{in}$ を選ぶ．

（d） 多項分布 $\mathrm{Mult}(\boldsymbol{\phi}_t)$ から単語 $\boldsymbol{w}_{in}$ を選ぶ．

DirTM のグラフィカルモデルを図 2 に示す．

3.3 DirTM の定式化

前節で仮定したグラフにおいて，各文書が属するノードからカテゴリ木の葉ノードに対応するノードへのパスに関する確率変数の集合を $\boldsymbol{X} = \{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_D\}$ と表す．ここで，$\boldsymbol{x}_i$ $(i \in \{1,\cdots,D\})$ は文書 i の各単語 n に関する，文書 i が割り当てられたノードからカテゴリ木の葉ノードに対応するノードに至るパスの集合を表す．このとき，DirTM を定式化したものを以下に示す．

$$\mathrm{P}(\boldsymbol{W},\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\xi}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}) \tag{1}$$

$$= \mathrm{P}(\boldsymbol{\phi}|\boldsymbol{\beta})\mathrm{P}(\boldsymbol{\xi}|\boldsymbol{\gamma})\mathrm{P}(\boldsymbol{W}|\boldsymbol{Z},\boldsymbol{\phi})\mathrm{P}(\boldsymbol{X}|\boldsymbol{\xi})\mathrm{P}(\boldsymbol{\theta}|\boldsymbol{\alpha})\mathrm{P}(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}) \tag{2}$$

$$= \left\{\prod_{t=1}^{T}\frac{\Gamma(\beta_\Sigma)}{\prod_w\Gamma(\beta_w)}\prod_{w=1}^{W}\phi_{tw}^{c(t,w)+\beta_w-1}\right\} \left[\left\{\prod_{v=1}^{V}\frac{\Gamma(\gamma_{v\Sigma})}{\prod_u\Gamma(\gamma_{vu})}\prod_{u=1}^{U}\xi_{vu}^{c(v,u)+\gamma_{vu}-1}\right\} \prod_{i=1}^{D}\left\{\prod_{\ell\in\mathcal{L}}\sum_{j\in\mathcal{D}_\ell}\frac{1}{|\mathcal{D}_\ell|}\frac{\Gamma(\alpha_{\ell\Sigma})}{\prod_t\Gamma(\alpha_{\ell t})}\prod_{t=1}^{T}\theta_{\ell jt}^{c(j,t)+\alpha_{\ell t}-1}\right\}^{\delta(u_i\notin\mathcal{L})}\right] \cdot\left[\left\{\prod_{i=1}^{D}\frac{\Gamma(\alpha_{\ell\Sigma})}{\prod_t\Gamma(\alpha_{\ell t})}\prod_{t=1}^{T}\theta_{it}^{c(i,t)+\alpha_{\ell t}-1}\right\}\right]^{\delta(u_i\in\mathcal{L})} \tag{3}$$

式 (3) の最初の $\{\cdot\}$ はトピック-単語分布に対応する項である．β_Σ は全語彙に対する β_w の総和である．$c(t,w)$ は語彙 w にトピック t が割り当てられる頻度である．

式 (3) の 1 つ目の $[\cdot]$ はカテゴリ木の内部ノードに対応する項である．u_i は文書 i が属するノードである．$\delta(\cdot)$ は括弧内の命題が真になるとき 1，そうでないとき 0 をとる関数である．$|\mathcal{D}_\ell|$ は $\mathcal{D}_\ell$ 中に含まれる文書の数である．γ_{vu} はノード対 (v,u) 間の辺 $v \to u$ に対する，ディリクレ事前分布のハイパーパラメータである．$\gamma_{v\Sigma}$ は v の全ての隣接ノードに対する γ の総和である．$c(v,u)$ は各内部ノードから各葉ノードへ至るパスの集合において，ノード対 (v,u) 間の辺 $v \to u$ が存在する数である．L をカテゴリ木の葉ノードの数とすると，$\alpha_{\ell t}$ $(\ell \in \{1,\cdots,L\},\ t \in \{1,\cdots,T\})$ は，各葉ノードごとに異なるディリクレ事前分布のハイパーパラメータである．$\alpha_{\ell\Sigma}$ は全トピックに対する $\alpha_{\ell t}$ の総和である．$c(j,t)$ は文書 j にトピック t が割り当てられる頻度である．

式 (3) の 2 つ目の $[\cdot]$ はカテゴリ木の葉ノードに対応する項である．$c(i,t)$ は文書 i にトピック t が割り当てられる頻度である．

式 (3) までの定式化は，3.2 におけるカテゴリ木の内部ノードの文書生成過程に基づいている．しかし，実際にこの仮定に基づくと計算コストが大きくなる．ここで，葉ノードのトピック分布を混合する際に，実際のカテゴリ木に存在する辺以外の辺についての確率が微小になることを利用して，内部ノードから葉ノードへのパスを，実際のカテゴリ木に存在するパスについてのみ考え，全ての葉ノード $\mathcal{L}$ についてではなく，文書 i が属する内部ノード v の子孫である葉ノード $\mathcal{L}_i$ のトピック分布に限って混合するという近似を行う．式 (3) に対してその近似を行うことにより，以下の式が得られる．

$$\mathrm{P}(\boldsymbol{W},\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\xi}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}) \tag{4}$$

$$\approx \left\{\prod_{t=1}^{T}\frac{\Gamma(\beta_\Sigma)}{\prod_w\Gamma(\beta_w)}\prod_{w=1}^{W}\phi_{tw}^{c(t,w)+\beta_w-1}\right\} \left[\left\{\prod_{v=1}^{V}\frac{\Gamma(\gamma_{v\Sigma})}{\prod_u\Gamma(\gamma_{vu})}\prod_{u=1}^{U}\xi_{vu}^{c(v,u)+\gamma_{vu}-1}\right\} \prod_{i=1}^{D}\left\{\prod_{\ell\in\mathcal{L}_i}\sum_{j\in\mathcal{D}_\ell}\frac{1}{|\mathcal{D}_\ell|}\frac{\Gamma(\alpha_{\ell\Sigma})}{\prod_t\Gamma(\alpha_{\ell t})}\prod_{t=1}^{T}\theta_{\ell jt}^{c(j,t)+\alpha_{\ell t}-1}\right\}^{\delta(u_i\notin\mathcal{L})}\right] \cdot\left[\left\{\prod_{i=1}^{D}\frac{\Gamma(\alpha_{\ell\Sigma})}{\prod_t\Gamma(\alpha_{\ell t})}\prod_{t=1}^{T}\theta_{it}^{c(i,t)+\alpha_{\ell t}-1}\right\}\right]^{\delta(u_i\in\mathcal{L})} \tag{5}$$

3.4 ギブスサンプリングによるモデル推定

ギブスサンプリングの際に用いる，文書 i の n 番目の単語に，葉ノードへのパス $x_{i\ell}$，トピック t が割り当たる完全条件付き確率 $\mathrm{P}(\boldsymbol{z}_{in}=t,\boldsymbol{x}_{in}=x_{i\ell}|\boldsymbol{W},\boldsymbol{X}_{-in},\boldsymbol{Z}_{-in},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma})$ を以下に示す．この式は式 (5) から導出できる．

$$\mathrm{P}(\boldsymbol{z}_{in}=t,\boldsymbol{x}_{in}=x_{i\ell}|\boldsymbol{W},\boldsymbol{X}_{-in},\boldsymbol{Z}_{-in},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}) \tag{6}$$

$$\propto \mathrm{P}(\boldsymbol{z}_{in}=t,\boldsymbol{w}_{in}=w,\boldsymbol{x}_{in}=x_{i\ell}|\boldsymbol{W}_{-in},\boldsymbol{X}_{-in},\boldsymbol{Z}_{-in},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}) \tag{7}$$

$$= \frac{c(t,w)-1+\delta(t\neq t')+\beta_w}{C_t-1+\delta(t\neq t')+\beta_\Sigma} \cdot \left[\left\{\sum_{j\in\mathcal{D}_\ell}\frac{1}{|\mathcal{D}_\ell|}\frac{c(j,t)+\alpha_{\ell t}}{N_j+\alpha_{\ell\Sigma}}\right\}\left\{\prod_{v\to u\in x_{i\ell}}\frac{c(v,u)+\gamma_{vu}}{P_v+\gamma_{v\Sigma}}\right\}\right]^{\delta(u_i\notin\mathcal{L})} \cdot \left[\frac{c(i,t)-1+\delta(t\neq t')+\alpha_{\ell t}}{N_i-1+\alpha_{\ell\Sigma}}\right]^{\delta(u_i\in\mathcal{L})} \tag{8}$$

ここで，$\ell\in\mathcal{L}_i$ である．P_v はカテゴリ木上の各内部ノードから各葉ノードへ至るパスの集合において，v を始点とする辺が存在する数である．t' は $\boldsymbol{z}_{in}$ に元々割り当たっていたトピックである．

4. 情報量基準によるモデル選択

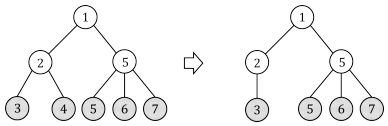

図 3　A example of merging child nodes of sub tree

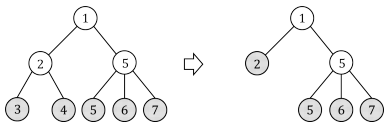

図 4　A example of merging all nodes of sub tree

DirTM において，葉ノードに属する文書集合の規模はそれぞれのノードによって様々である．もし，葉ノードに属する文書集合の規模が十分大きくない場合に，ハイパーパラメータの推定において過適合が起こるという問題がある．これは，モデルのトピック数と葉ノードに属する文書集合の規模がうまく合致していないことによる．この問題が起きると，葉ノードにおいて正確なトピック分布の推定が行われず，モデル推定精度の不安定さにつながるおそれがある．

この問題を解決するために，情報量規準を用いて各ノードにおいてモデル選択を行う手法を提案する．情報量規準はモデルの良さを測るための指標であり，代表的なものに赤池情報量規準 (AIC: Akaike Information Criterion) がある [2], [3]．AIC を以下に示す．ここで，$\log l$ はモデルの最大対数尤度，p はモデルのパラメータ数を表す．

$$\mathrm{AIC} = -2\log l + 2p \tag{9}$$

モデルの訓練データに対する過適合問題を解決するために，情報量規準を用いる．式 (9) においてパラメータ数を増加させるとき，右辺第 1 項が単調減少するのに対し，右辺第 2 項のパラメータ数は単調増加するペナルティ項となり，モデルの過適合を防ぐ．すなわち，AIC の値が最小となるときのパラメータ数が，そのモデルの最も良いパラメータ数である．このことから，AIC の値が小さいモデルほど，モデルとして良いといえる．

DirTM におけるモデル選択について以下に述べる．DirTM において，パラメータ数 p はハイパーパラメータ $\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}$ の数の合計である．ここで，$\boldsymbol{\alpha}$ はトピック数 × カテゴリ木の葉の数，$\boldsymbol{\beta}$ は語彙数，$\boldsymbol{\gamma}$ は根を除くノードの総数，すなわち辺の総数だけ存在する．よって，パラメータ数 p は以下のように書ける．ここで，L はカテゴリ木の葉の数，W は語彙数，E はカテゴリ木の辺の総数である．

$$p = TL + W + E \tag{10}$$

よって，DirTM における AIC は以下のように書ける．

$$\mathrm{AIC}_{\mathrm{DirTM}} = -2\log l + 2(TL + W + E) \tag{11}$$

まず，ギブスサンプリングによるモデル推定後の訓練文書に対する対数尤度を最大対数尤度として用いて，元のモデルの AIC を式 (11) に従って計算する．次に，内部ノードを根とする部分木を 1 つランダムに選び，部分木中のノードを併合して 1 つのノードと見なす．つまり，その部分木全体を 1 つの葉ノードと見て，部分木の根以外のノードに属していた文書を部分木の根に属するように変更する．このときに，カテゴリ木の葉の数が減少するので，式 (11) におけるパラメータ数が変化する．その上で，葉ノードに関する $\boldsymbol{\alpha}$ の推定，訓練文書に対する対数尤度の計算を行い，それらを元に式 (11) に従って新しい AIC の値を計算する．元のモデルと新たに構成したカテゴリ木によるモデルの AIC の値を比較し，新しいモデルの方が AIC が小さい場合，新しいモデルの方が適切であると判断しモデルを更新する．モデル更新時にカテゴリ木を変形する方法として，以下の 2 通りの方法を考える．

（1） 着目している部分木の葉ノードのみを併合する (DirTM-AIC1).

（2） 着目している部分木の全ノードを併合する (DirTM-AIC2).

図 3，4 を用いて具体的に説明する．今，図においてカテゴリ 2 を根とする部分木に着目したとする．(1) の方法の場合，モデル更新時に，図 3 のように，部分木の葉ノードのみを併合し，葉ノードに属していた文書は，併合されて新たにできたカテゴリ 3 に全て属するように変更する．そして，部分木中の内部ノードと新たにできた葉ノードそれぞれに属する文書についてのみ，ギブスサンプリングを用いてトピック分布を推定し直す．(2) の方法の場合，図 4 のように，部分木の全てのノードを併合し，1 つの葉ノードと見なす．部分木の各ノードに属していた文書は，全て新たにできたカテゴリ 2 に属するように変更する．そして，新たにできた葉ノードであるカテゴリに属する文書についてのみ，ギブスサンプリングを用いてトピック分布を推定し直す．以上のどちらの方法を用いる場合も，この操作を部分木の根をなす内部ノード全てに対して行う．

以上の手法を用いて，葉ノードにおけるハイパーパラメータの過適合を防ぎ，モデル推定精度の向上を目指す．

5. 実　　験

提案手法である DirTM のモデル選択付き学習の有用性を示すため，実際にギブスサンプリングに基づくモデル推定，モデル選択の実験を行う．

5.1 MEDLINE データセット

表 1　Summary of a sub set (B02) of MEDLINE data set

number of documents	4971
number of word types	9019
number of word tokens	391374
number of MeSH	53

生物医学系のデータベースである MEDLINE のデータセットをモデル推定に用いる．MEDLINE 中の各文書には，MeSH (Medical Subject Headings) と呼ばれる生物医学用語が 10 から 15 個付与されている．また，MeSH は階層構造をなしており，下位の層ほどその意味が具体的，限定的になっていく．MeSH は文献の内容を説明しており，一種のカテゴリと見なすことができる．本実験では，この MeSH をカテゴリとして用いることとする．

本実験では，2009 年の MEDLINE データセットから，MeSH ワード algae (藻類) を根とする部分木であるカテゴリ階層を取り出し，そのカテゴリ階層中のカテゴリが付与された文書集合を用いて，モデル推定を行う．MEDLINE 中の各文書には 10 から 15 個の MeSH が付与されているが，DirTM では各文書について 1 つのカテゴリが付与されていることを前提としているので，文書が複数の MeSH を持つときは，階層構造中の最下層の MeSH を 1 つ残し，それ以外の MeSH は取り除く．また，情報検索システム INQUERY [6] で用いられた 418 種類のストップワードを，データセットから予め除去する．さらに，10 単語以下からなる文書，5 文書以内にしか現れない単語を予め除去する．以上の準備を行った後のデータセットの概要を表 1 に示す．

5.2 モデル推定

5.2.1 テストセット対数尤度

モデル推定の際に，評価尺度としてテストセット対数尤度を用いる．テストセット対数尤度が大きいほど，モデルの精度が高いことを意味している．テストセット対数尤度は，トピックモデルなどの統計的言語モデルの評価尺度としてよく知られている．DirTM におけるテストセット尤度 $\mathrm{P}(\boldsymbol{w}_{\mathrm{test}})$ は以下の式で表される．

$$\mathrm{P}(\boldsymbol{w}_{\mathrm{test}}) \tag{12}$$

$$= \prod_i \prod_w \sum_t \frac{c(t,w) - 1 + \beta_w}{C_t - 1 + \beta_\Sigma} \cdot \left[\left\{ \sum_{j \in \mathcal{D}_\ell} \frac{1}{|\mathcal{D}_\ell|} \frac{c(j,t) + \alpha_{\ell t}}{N_j + \alpha_{\ell\Sigma}} \right\} \left\{ \prod_{v \to u \in x_{i\ell}} \frac{c(v,u) + \gamma_{vu}}{P_v + \gamma_{v\Sigma}} \right\} \right]^{\delta(u_i \notin \mathcal{L})} \cdot \left[\frac{c(i,t) - 1 + \alpha_{\ell t}}{N_i + \alpha_{\ell\Sigma}} \right]^{\delta(u_i \in \mathcal{L})} \tag{13}$$

テストセット対数尤度は (13) 式の対数をとったものである．

本実験では，データセット中の 90%の単語を訓練セット，10%をテストセットに分割する．訓練セットを使ってモデルの推定を行い，テストセットを使ってそのモデルのテストセット対数尤度を測定することで，モデルの精度を評価する．

5.2.2 実 験 設 定

実験において，式 (8) に従ってギブスサンプリングすることによりモデルを推定する．トピック数 T は $T = 25, 50, 75$ の 3 通りとする．また，ハイパーパラメータ $\boldsymbol{\alpha}$ は初期値を $\alpha_{\ell t} = 50/T$ とし，ギブスサンプリング 5 回毎に，不動点反復法により 5 回反復して推定する．このとき，ギブスサンプリングの最初の 10 回では $\boldsymbol{\alpha}$ の推定は行わない．ハイパーパラメータ $\boldsymbol{\beta}$ は $\beta_w = 0.1$ で固定する．ハイパーパラメータ $\boldsymbol{\gamma}$ は，どのノード対 (v, u) に対しても同じとし，$\gamma = 0.01, 1, 100$ の 3 通りで実験する．

ギブスサンプリングの収束条件を設定する．推定しているモデルのテストセット対数尤度をギブスサンプリング 10 回毎に測定し，その変化率が ±0.1% 以内に収まるとき，ギブスサンプリングを打ち切る．

5.2.3 実 験 結 果

表 2　Test set log likelihood of LDA

T	total	leaf node	internal node
25	-7.171	-7.150	-7.263
50	-7.153	-7.137	-7.224
75	-7.128	-7.109	-7.177

表 3　Test set log likelihood of DirTM ($\gamma = 0.01$)

T	total	leaf node	internal node
25	-7.265	-7.256	-7.304
50	-7.071	-7.165	-6.644
75	-6.945	-7.101	-6.241

表 4　Test set log likelihood of DirTM ($\gamma = 1$)

T	total	leaf node	internal node
25	-7.271	-7.265	-7.295
50	-7.070	-7.167	-6.629
75	-6.958	-7.116	-6.244

表 5　Test set log likelihood of DirTM ($\gamma = 100$)

T	total	leaf node	internal node
25	-7.265	-7.259	-7.318
50	-7.080	-7.175	-6.652
75	-6.967	-7.129	-6.235

モデル推定を行った LDA と DirTM の各単語ごとのテストセット対数尤度を，表 2 から表 5 に示す．有効数字は 4 桁とする．ここでは，テストセット中の文書全体に対するテストセット対数尤度，葉ノード，内部ノードに属する文書のみを見たと

きのテストセット対数尤度の 3 つを示している．表を見て分かるように，トピック数が 50，75 のときに DirTM が LDA に対してより良い値をとっている．葉ノードに属する文書に対する推定精度をほぼ維持しつつ，さらに，内部ノードに属する文書に対して，LDA と比較して推定精度が改善されていることがわかる．すなわち，DirTM がカテゴリ階層における葉ノードと内部ノードの関係をより捉えたモデル化を行うことができているといえる．

また，ハイパーパラメータ $\boldsymbol{\gamma}$ の値によってテストセット対数尤度を比較すると，γ が小さいほど良い値を取る傾向にあった．

5.3 モデル選択

推定したモデルに対して，提案手法によるモデル選択を行う．

5.3.1 実験設定

実験では，DirTM を推定し，簡単のためカテゴリ木の最も深いノードの 1 つから幅優先探索により部分木を選び，AIC によってモデルの評価を行う．トピック数は $T = 25, 50, 75, 100, 125$ の 5 通りとする．また，ハイパーパラメータ $\gamma = 0.01$ に固定している．モデルを更新する場合には，4. で述べたように，DirTM-AIC1 と DirTM-AIC2 について実験を行い，モデル選択を伴わない DirTM および LDA をベースラインとして比較を行う．DirTM-AIC1 は，着目している部分木の葉ノードのみを併合し，その部分木中の全ノードに属する文書についてのみ，式 (8) に従ってギブスサンプリングでトピック分布を推定し直す．DirTM-AIC2 は，着目している部分木の全ノードを併合し，併合された 1 つのノードを葉ノードと見なして，それに属する文書についてのみ，式 (8) に従ってギブスサンプリングでトピック分布を推定し直す．また，ハイパーパラメータ $\boldsymbol{\alpha}$ は初期値をモデル選択において AIC を計算する際に推定したものを用いて，ギブスサンプリング 5 回毎に，不動点反復法により 5 回反復して推定する．

5.3.2 実験結果

表 6　Test set log likelihood of DirTM-AIC1 ($\gamma = 0.01$)

T	total	leaf node	internal node
25	-7.170	-7.263	-6.735
50	-7.039	-7.165	-6.470
75	-6.914	-7.101	-6.069
100	-6.902	-7.108	-5.973
125	-6.849	-7.091	-5.757

表 7　Test set log likelihood of DirTM-AIC2 ($\gamma = 0.01$)

T	total	leaf node	internal node
25	-7.247	-7.298	-6.968
50	-7.045	-7.178	-6.318
75	-6.952	-7.142	-5.913
100	-6.880	-7.106	-5.760
125	-6.843	-7.108	-5.527

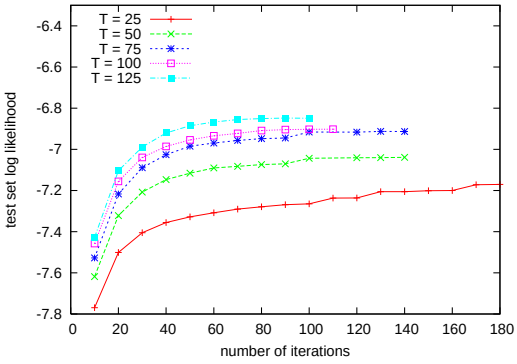

図 5　Test set log likelihood of DirTM-AIC1 per 10 times of iteration

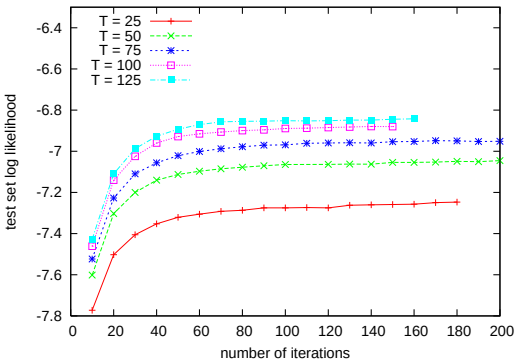

図 6　Test set log likelihood of DirTM-AIC2 per 10 times of iteration

モデル選択後における，各単語ごとのテストセット対数尤度の測定結果を表 6，7 に示す．有効数字は 4 桁とする．表 6 が，DirTM-AIC1，すなわちモデル更新時に部分木の葉ノードのみを併合する方法でモデル選択を行った結果であり，表 7 が，DirTM-AIC2，すなわちモデル更新時に部分木のノード全てを併合する方法でモデル選択を行った結果である．表 3 と比較すると，モデル選択を行うことによって，全体としてテストセット対数尤度が改善されていることがわかる．

また，ギブスサンプリングの反復 10 回ごとの，各単語ごとのテストセット対数尤度の測定結果を図 5，6 に示す．図 5 が，DirTM-AIC1，すなわちモデル更新時に部分木の葉ノードのみを併合する方法でモデル選択を行った結果である．$T = 25$ において，反復回数の初期は DirTM と等価であるが，反復回数 110 回で収束が判定され，その後は順次選択された部分木に対してモデル選択によるギブスサンプリングを行なっている．該当する曲線において反復回数 100 回で段ができているのはこのためである．同様に，$T = 50, 75, 100$ は反復回数 100 回から，$T = 125$ は反復回数 90 回からモデル選択によるギブスサンプリングを行なっている．図 6 が，DirTM-AIC2，すなわちモデル更新時に部分木のノード全てを併合する方法でモデル選択を行った結果である．図 5 と同様に，$T = 50, 100$ で反復回数 100 回からモデル選択によるギブスサンプリングを行なっている．同様に，$T = 25, 125$ は反復回数 90 回から，$T = 75$ は

反復回数 110 回からモデル選択によるギブスサンプリングを行なっている．図 5，6 を見ると，DirTM-AIC1，DirTM-AIC2 の両手法とも，モデル選択を行ったときに，テストセット対数尤度が改善され，さらに収束していることが分かる．

以上のように，モデル選択付き DirTM の推定の際に内部ノードにおいて値が改善され，それによって全体のテストセット対数尤度が改善されている．これは部分木における葉ノードの併合を行いトピック分布を推定し直すことによって，その葉ノードにおけるハイパーパラメータ $\boldsymbol{\alpha}$ の過適合が改善され，内部ノードのトピック分布を推定する際に影響を与えていると考えられる．また，葉ノードについてはおおむね精度に変化は見られない．これは，ハイパーパラメータ $\boldsymbol{\alpha}$ の過適合が改善されるものの，モデル選択を行うことによって葉ノードに属する文書集合が大きくなり推定精度が落ちるため，トピック分布の推定精度はほぼ横ばいのままとなるためと考えられる．

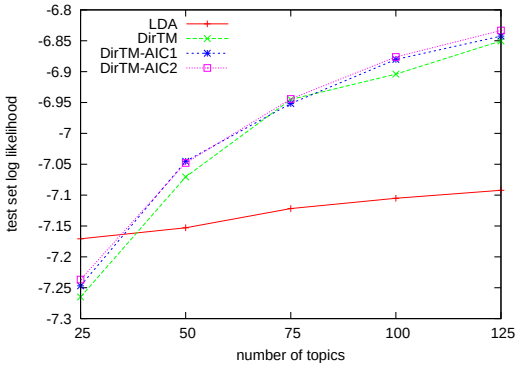

図 7　Test set log likelihood of all model

また，図 7 に，$T = 25, 50, 75, 100, 125$ における LDA，DirTM，DirTM-AIC1，DirTM-AIC2 のそれぞれのテストセット対数尤度の収束状態を比較したものを示す．$T \geq 50$ において提案手法である DirTM のモデル選択付き学習が，ベースラインである LDA，モデル選択なし DirTM より優れた結果を残していることが分かる．また，4. で提案した 2 つの方法を比較すると，テストセット対数尤度にほとんど差は見られない．

6. おわりに

本論文では，従来のトピックモデルで直接取り扱うことのなかったカテゴリ階層に対し，そのカテゴリをモデル化の際に考慮した確率的トピックモデルである DirTM を提案した．また，モデル推定の際にカテゴリ木の構造に着目してモデル選択を行う手法を提案した．また，ギブスサンプリングによる推定後にモデル選択を行ったモデルのテストセット対数尤度を測定し，代表的なトピックモデルである LDA と比較して，その有効性を示した．

今後の課題としては，カテゴリ階層が付与された様々なデータに対する評価，実験結果に対する詳細な分析，応用タスクへの適用の有効性を評価が挙げられる．また，本研究では，カテゴリ木の構造に着目し，カテゴリ木の変形を行うことで，あるトピックでの最適なモデルを選択するという手法をとった．それとは別のアプローチにとして，階層ディリクレ過程 (HDP: Hierarchical Dirichlet Process) [14] によるトピック数の自動的な決定を行う，という方法も考えられる．

謝　辞

本研究の一部は，科学研究費補助金基盤研究 (B) (20300038, 23300039) および国立情報学研究所共同研究による．

文　献

[1] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed Membership Stochastic Blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.

[2] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, pp. 267–281, 1973.

[3] 赤池, 甘利, 北川, 樺島, 下平. 赤池情報量規準 AIC. 共立出版, 2007.

[4] D. M. Blei and J. D. McAuliffe. Supervised topic models. *Advances in Neural Information Processing Systems*, 21, 2007.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[6] J. P. Callan, W. B. Croft, and S. M. Harding. The INQUERY Retrieval System. In *Proceedings of the 3rd International Conference on Database and Expert Systems Applications*, pp. 78–83, 1992.

[7] L. Fei-Fei and P. Perona. A Bayesian Hierarchical Model for Learning Natural Scene Categories. In *Computer Vision and Pattern Recognition*, Vol. 2, pp. 524–531, 2005.

[8] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceeding of the National Academy of Sciences of the United States of America*, 101:5228–5235, 2004.

[9] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative Learning for Dimensionality Reduction and Classification. *Advances in Neural Information Processing Systems*, 21, 2008.

[10] W. Li and A. McCallum. Pachinko allocation: DAG-structured mixture models of topic correlations. In *ICML '06 Proceedings of the 23rd international conference on Machine learning*, 2006.

[11] T. P. Minka. Estimating a Dirichlet distribution. Technical report, Microsoft Research, 2003.

[12] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP '09 Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Vol. 1, 2009.

[13] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, 2004.

[14] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101:1566–1581, 2006.

[15] H. M. Wallach, D. Mimno, and A. McCallum. Rethinking LDA: Why Priors Matter. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems*, 2010.